An Audiobooks-based Approach for Creating a Speech Corpus for Acoustic Models

Salvatore Michele Biondi, Vincenzo Catania, Ylenia Cilano, Raffaele Di Natale, Antonio Rosario Intilisano,
Giuseppe Monteleone, Daniela Panno
Dipartimento di Ingegneria Elettrica Elettronica e Informatica
University of Catania
Catania, Italy
{salvo.biondi, vincenzo.catania, ylenia.cilano, raffaele.dinatale, aintilis, giuseppe.monteleone, daniela.panno}@dieei.unict.it

***Abstract*—The limited availability of a valid speech corpus is one of the major problems affecting the design of speech recognition acoustic models. As a matter of fact, large amounts of manually-transcribed data is necessary in order to build a valid acoustic model. Nevertheless, obtaining large datasets is generally both time- and resource- consuming as it requires a continuous supervision of the entire building process. Speech corpora can be used to generate an acoustic model, however a large part of these are not suitable or freely available. This paper aims at showing the use of audiobooks as databases for creating speech corpora. An automatic algorithm that processes audiobooks for building speech corpora is proposed. This method allows to replace traditional manual transcription of audio- recordings and to automatically obtain a phonetic dictionary. An Italian acoustic and linguistic model was generated as use-case to test the effectiveness of the proposed procedure.**

Keywords-Automatic Speech Recognition; Audio Databases; Audiobook; Speech processing.

I. INTRODUCTION

Automatic Speech Recognition (ASR) maps an acoustic signal into a sequence of phonemes or words (discrete entities). A typical ASR process performs the decoding of input speech data by means of signal processing techniques and acoustic or language models. Acoustic models refer to the representation of acoustics and phonetics, while language models describe the words that are recognized by the ASR process.

The complexity of natural language makes its use for human – computer interaction a difficult task to perform [1]. For example, a specific context may influence the generation of phonemes or speech signal can vary significantly according to speaker sex, style, speed and can be affected by noise.

Speech recognition engines require statistical representation of each of the distinct sounds that makes up a word (Acoustic model). This model is created by a training process that compiles recordings and their transcriptions into a statistical representation of the sounds for each word. Therefore, large amounts of transcribed recordings are generally required.

Obtaining these data is an expensive and time-consuming task: several hours of recordings are necessary, recorded in a quiet environment and by different voices. Finding all these data with the relative transcription is not trivial.

An alternative to build from scratch a database with audio recordings and manual transcriptions is the use of audiobooks.

This paper is organized as follows: the prior works are described in Section II, Section III depicts the novel approach, Section IV the procedure of training acoustic models, Section V describes the algorithm developed to obtain the phonetic transcriptions, Section VI shows the results; finally, in Section VII, we draw conclusions and possible future works.

II. RELATED WORK

Different approaches are available to generate an audio corpora database to be used for creating an acoustic model: a first approach [2] that uses speakers with different age and gender for providing recordings related to a predefined text; a second approach that uses available audio sources (radio broadcast, news broadcasts) and the corresponding text transcriptions [3][4]. However, the first method is accurate but time- and resource-consuming, while the second one is easy to develop but it often provides incomplete results because most of its audio data sources have no corresponding accurate word transcriptions.

For example, a complete commercial recording database of Italian language records already exists and is called APASCI [5]. This is an Italian speech database designed for researching on acoustic modeling. The process to create a database using these features from scratch requires much time.

Another available solution is Voxforge [6], a free project that collects corpora of spoken speeches in several languages thanks to the collaboration of users who provide their voices. All audio files are available under the General Public License (GPL) and allow obtaining acoustic models for many speech recognition engines such as Carnegie Mellon University (CMU) Sphinx [7] and Hidden Markov Model Toolkit (HTK) [8]. Sphinx does not provide an Italian acoustic model, thus, it has to be created with a suitable audio database. Beside Voxforge, no Italian free audio databases, which work with the aforementioned engine, were created.

At the time of writing this paper Voxforge contained about 11 hours of recordings for Italian language. Such a small amount of records does not allow to obtain a good acoustic model using Sphinx, because to create a new model 50 hours of audio recordings of 200 different speakers are required [9]. So, the need of a specific audio database for creating an Italian language acoustic model with Sphinx has arisen.

III. A NOVEL APPROACH

As described above, the acoustic model is a building block of an ASR system and it can manage a specific language only if an acoustic model for that language is available.

In this paper, a novel approach for creating an acoustic and linguistic model for Sphinx is described. It provides a method to obtain in a simple and fast way many transcribed recordings from audiobooks. Audiobooks are a valid and reliable source to create acoustic models because they are recorded in an echoic chamber by different voices and the text of the audiobook is available. Audiobooks are a good and free statistical basis in terms of cost / time, but the problem is that, in general, the training for creating an acoustic model requires small audio files in terms of duration (for example the optimal length for Sphinx is not less than 5 seconds and not more than 30 seconds [9]). Because audiobooks provides audio files that are too long for Sphinx, a method to split audiobooks and to associate to each part obtained the corresponding transcription is necessary. The tool HTK was used for this purpose. To solve the problem of the creation of acoustic models, we need to provide the Italian phonetic transcription of each word in audio files. An algorithm that creates phonetic transcriptions has been developed For the training of the Italian acoustic model we used SphinxTrain [9] (it is the tool used for the training of an acoustic model for Sphinx).

This method can be applied for the creation of acoustic models in several languages.

IV. TRAINING BY AUDIOBOOK

Audiobooks are usually available like a single long audio file with the corresponding text transcription. For our research, we created an automatic method to develop a training set for SphinxTrain from audiobooks. So, this method splits a single audiobook in several little audio files with the corresponding text transcription.

These audio files make the Audio Database required from sphinx to extract statistic from the speech.

HTK toolkit [10] was used to split the audiobooks. HTK [11] requires a little acoustic model to perform the work described above. This acoustic model was developed by using VoxForge training through a set of collected transcribed speech corpora. During tests, VoxForge held a database with about 11 hours recording with related text transcriptions. Usually, an audio database with 11 hours of recording is not enough to create an efficient acoustic model, but it was sufficient to use in our automatic split method through HTK.

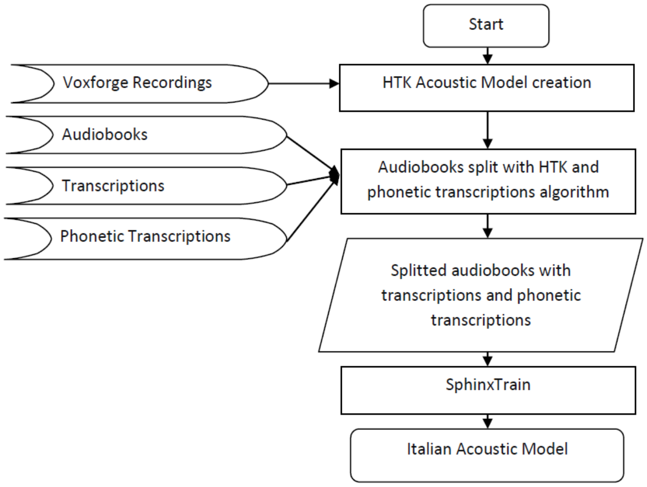
Start
Voxforge Recordings
HTK Acoustic Model creation
Audiobooks
Transcriptions
Phonetic Transcriptions
Audiobooks split with HTK and phonetic transcriptions algorithm
Splitted audiobooks with transcriptions and phonetic transcriptions
SphinxTrain
Italian Acoustic Model

Figure. 1. Procedure of training

Our test shows that this database is a valid choice for our task. In fact, we use this first model to split the selected audiobooks because it allows to align the audio to its transcription.

Next, the phonetic transcriptions of the words must be provided: for this purpose, the algorithm described in Section 5 has been developed. Finally, audio recordings obtained by the split audiobooks and their phonetic transcriptions are input for SphinxTrain that produces an Italian acoustic model. Figure 1 shows the whole procedure.

V. PHONETIC DICTIONARY GENERATION

The training function must be configured with the sound units, the corresponding transcriptions and the phonetic dictionary. It maps every word into a sequence of sound units (phonetic transcription). To derive the sequence of sound units, the phonetic transcriptions are associated with each word:

ABACO	a b a k o
ABACHI	a b a k I

In the example above, the word ABACO is the Italian for “Abacus”, while ABACHI is its plural. At their right, we show their corresponding phonetic translation.

Initially, it uses an existing dictionary that consists of a customized version of the Festival [12] lexicon (brackets and number were removed). It includes more than 440,000 words, but it is not complete enough to perform the phonetic transcriptions of all the words present in the split audio files. Accents are important for Italian language because the meaning of the word may change based on their position.

So, an algorithm to derive the missing phonetic transcriptions is necessary. The developed algorithm is able to derive two different phonetic transcriptions. The input of the algorithm consists of a list of words not found in the Festival lexicon and for each word it performs the following steps:
- [optional] an online search of the word in an Italian dictionary-database [13]: if the word is found in the dictionary the position of the accent is obtained. In fact, if the word is a lemma (or entry-word) missing in the phonetic dictionary, the algorithm generates the phonetic transcription with the correct accent. This step is optional since calling the web service takes a considerable time. Furthermore, grandidizionari.it service does not contain all the necessary words, so:
- If the accent was not obtained, the lemma from which the word is derived is looked up in the dictionary of Morph-IT! [14], that is a lexicon of inflected forms with their lemma and morphological features. Output of this module is a tuple composed by:

1. Form
2. Lemma
3. Features

- At this point, the substring with which the word ends is compared with a list of available desinences written according to the following format:

```
vowel_or_consonant+desinence,category,
part_of_speech,phonetical_transcription
```

-`vowel_or_consonant` indicates if "`desinence`" must be preceded by a vowel, a consonant, or both. This parameter is optional and can take 3 values: V (vowel), C (consonant), VC (vowel plus consonant).
- `desinence` is the analyzed desinence, that is the string compared with the substring with which the word ends. The desinences have been obtained from [15]. For each desinence, its inflected versions are obtained by analyzing the `category`.
- `category` is used to get the inflections of each desinence. A letter represents each category. There is a list of categories in which each category is associated with some characteristics. Each characteristic is written according to the following format:

$$\texttt{category: } s_{1,1}\text{-}pt_{1,1} > si_{1,1}\text{-}pti_{1,1},\ldots,\ si_{1,n}\text{-}pti_{1,n};\ \ldots\ ;\ s_{k,1}\text{-}pt_{k,1} > si_{k,1}\text{-}pti_{k,1},\ldots,\ si_{k,n}\text{-}pti_{k,n};$$

(where s=substring, pt=phonetic transcription, i=inflected).

In general, if a desinence belongs to a certain category and it ends in $s_{k,1}$, to obtain the inflected form, $s_{k,1}$ is removed from the desinence and it is replaced with $si_{k,1}$, $pt_{k,1}$ is removed from its phonetic transcription and it is replaced with $pti_{k,1}$. This procedure is applied for all eventual n inflections.

For example, consider the "D" category that is written as:

```
D:gia-dZ i! a>gie-dZ i! e;cia-tS i! a>cie-tS i! e;
```

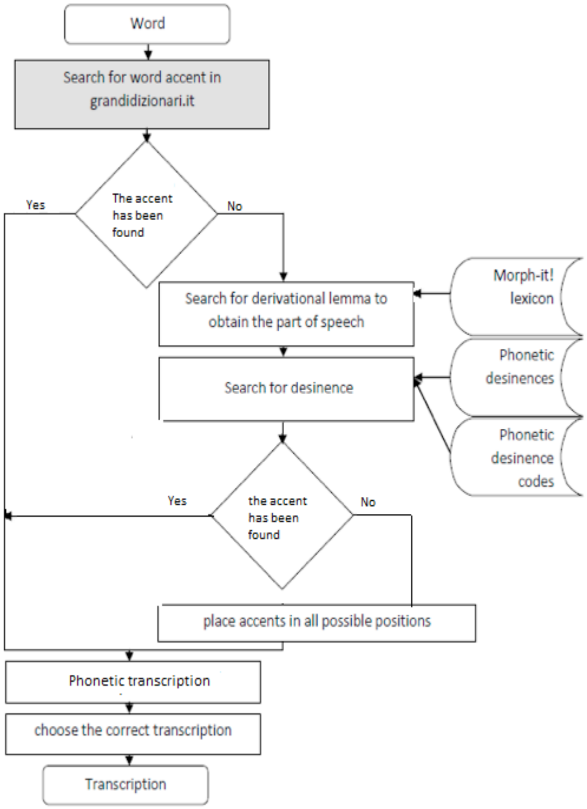
Word
Search for word accent in grandidizionari.it
The accent has been found
Yes
No
Search for derivational lemma to obtain the part of speech
Morph-it! lexicon
Search for desinence
Phonetic desinences
Phonetic desinence codes
the accent has been found
Yes
No
place accents in all possible positions
Phonetic transcription
choose the correct transcription
Transcription

Figure. 2. Algorithm for phonetic transcriptions, in gray the optional step

This means that if a desinence belongs to the category D, if it ends in "-gia" and the correspondent phonetic transcription ends in "-dZ i! a", the inflected form is obtained by removing "-gia" and adding "-gie " in the desinence and by eliminating " dZ-i! a" and adding "-dZ i! e" in the phonetic transcription.

There is also the "I" category for desinences that do not have inflected forms: `I:;`

- `part_of_speech` indicates the part of speech and is used because some equal desinences endings have different accents emphasis depending on the part of speech. The used parts of speech are: V for verbs, N for nouns and adjectives, A for other;
- `phonetical_`transcription is the phonetic transcription of the desinence.

If no desinence has been identified and the word does not have an accent, a set of strings is created. This set includes all possible inflected forms for the analyzed word. Transcriptions are created for each of these forms and the user can choose the correct one. The entire procedure is shown in Figure 2.

For the Italian language, the Italian phonemes are used [16][17] with the following simplification: close-mid front unrounded vowels and open-mid front unrounded vowels are mapped as a single phoneme “e”. The same simplification is applied to the close-mid back rounded vowels and the open-mid back rounded vowels, mapped in a single phoneme “o”.

The algorithm provides the phonetic transcription of a word both with accents and without (setting by the software), and the user can choose which version to take into account. The word to be phonetically transcribed is analyzed character by character. Optionally, the software cannot take into account the position of accent and therefore does not generate a phonetic transcription that includes the accent, skipping the entire branch "no" in Figure 2 of the first if statement (including grandidizionari.it service).

VI. Experimental Results

The goal of this test was not to validate the Sphinx acoustic model we developed for the Italian language, since no other models were available for comparison. Instead, the purpose of the test was to demonstrate that the approach based on audiobooks can deliver a valid acoustic model. We expect that increasing the size of the training set should provide higher performance.

A first acoustic model with HTK has been created by using Italian VoxForge database. The splitting algorithm has been applied to the free audiobooks [18]. We obtained about 34 hours of split recordings, 56% of them being spoken by female voices. Audiobooks were read by six different male voices and two different female voices.

The automatic phonetic generation gave the same results of querying the “grandidizionari.it” web service. In addition, we obtained phonetic translation for words not provided by the online dictionary.

Finally, a linguistic model was obtained using the transcriptions of audiobooks. To test the performance of the acoustic model two different speakers (a male and a female) pronounced a list of 40 words taken from the vocabulary of the linguistic model. Table 1 shows the results:

TABLE I. Results

	speaker 1 (female)	speaker 2 (male)
recognized words %	70%	77.50%

The acoustic model was trained with two female and six men voices. Although female voices are 56% compared to the total, the male voice is better recognized than the female one. The results show that the variety of voices affects the quality of the result [19]. In addition, we noted that audio recordings that did not match with corresponding words are recognized as very similar words from the phonetic point of view (for example, the word “velocemente” is recognized as “velatamente”).

VII. Conclusions and future work

This paper shows an automatic method to obtain recordings, transcriptions and phonetic transcriptions in order to create an acoustic and linguistic model. Our goal was to investigate the possibility of using a set of free audiobooks for generating a dataset as a complete database of ad hoc audio recordings. Then, an Italian acoustic model has been created by SphinxTrain, actually not available in [20]. Currently, the performed tests are based on approximately 30 hours of recordings.

In order to obtain an enhanced Italian acoustic model, a selection of at least 50 hours of audiobooks recordings is required. The audiobooks recordings must be selected by different voices, ages, genders of the speaker or the topic of the audiobook. In total, about 200 different speakers are needed.

Acknowledgements

Authors were in part supported by the Sicily Region grants PROGETTO POR 4.1.1.1: "Rammar Sistema Cibemetico programmabile d’interfacce a interazione verbale".

References

[1] M. Forsberg, "Why is speech recognition difficult?" Technical Report from Department of Computing Science Chalmers University of Technology, Sweden, 2003.

[2] VoxForge, http://www.voxforge.org (last visited March 5, 2014).

[3] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training", Computer Speech & Language, Spoken Language Processing Group, CNRS-LIMSI, Orsay, France, Volume 16 Issue 1, 2002, pp. 115-129.

[4] C. Cieri, D. Graff, and M. Liberman, “The TDT-2 Text and Speech Corpus” in In Proceedings of Darpa Broadcast News Workshop, 1999, pp. 57–60.

[5] APASCI, http://catalog.elra.info/product_info.php?products_id=168 (last visited February 25, 2014).

[6] VoxForge repository, http://www.repository.voxforge1.org/downloads/it/Trunk/Audio/ (last visited March 5, 2014).

[7] K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system, "IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 38, No. 1, January, 1990, pp. 35 - 45.

[8] HTK Documentation, http://htk.eng.cam.ac.uk/docs/docs.shtml (last visited February 25, 2014).

[9] Training Acoustic Model for CMUSphinx, http://cmusphinx.sourceforge.net/wiki/tutorialam (last visited March 3, 2014).

[10] HTK Toolkit and tutorial, http://www.voxforge.org/home/dev/autoaudioseg (last visited March 3, 2014).

[11] HTK JuliusTutorial, http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial (last visited March 4, 2014).

[12] Festival website, http://festvox.org/ (last visited March 3, 2014).

[13] Grandizionari online service, http://www.grandidizionari.it/Dizionario_Italiano.aspx?idD=1 (last visited March 3, 2014).

[14] E. Zanchetta and M. Baroni, "Morph-it! A free corpus-based morphological resource for the Italian language" , proceedings of Corpus Linguistics 2005, University of Birmingham, Birmingham, UK.

[15] Desinences list, http://www.dipionline.it/guida/docs/DiPI_Terminazioni_Desinenze.pdf (last visited March 3, 2014).

[16] F. Albano Leoni and P. Maturi, “Manuale di fonetica” ,2002, Roma, Carocci.

[17] P. Maturi, “I suoni delle lingue, i suoni dell'italiano”, 2006, Bologna, Il Mulino.

[18] Free audiobooks repository from librivox, http://librivox.org/ (last visited March 3, 2014).

[19] M. Gerosa, D. Giuliani, and S. Narayanan: "Acoustic analysis and automatic recognition of spontaneous children’s speech", In Interspeech-2006, paper 1082-Wed2CaP.9.

[20] SPHINX free language model repository, http://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/ (last visited March 3, 2014).

ISBN: 978-1-61208-372-8